# Theoretical and Computational Solutions for Phraseological Lexicography

**Richard Almind, Henning Bergenholtz, Vibeke Vrang (Aarhus)**

**Abstract**

THE DANISH IDIOM DICTIONARY has been criticized by Ken Farø to use a less than optimal theory in the description of what a phraseme really is. A closer examination of that theory, however, shows that this is not the case due to the hitherto neglected role of the common, non-academic user who has little need for an in-depth classification. The current article describes how THE DANISH IDIOM DICTIONARY puts the simplified theory into use and how a new and improved version currently under way improves on the theory by focusing on the user's needs, thereby reducing the necessity for complicated classifications. THE DANISH IDIOM DICTIONARY including its search capabilities is described in full.

## 1 Phraseological Online Dictionaries

The exact number of phraseological online dictionaries in existence and the languages for which they are made cannot be established with certainty, but to our knowledge the number is not impressive. Most online dictionaries are designed in exactly the same way as printed dictionaries, often in order to promote the printed version, cf. GOENGLISH.COM (2005), which gives an example for each proverb, but no explanation, e. g.

**A Bird In The Hand Is Worth Two In The Bush**

"Dan has asked me to go to a party with him. What if my boyfriend finds out?"
Reply: "Don't go. A bird in the hand is worth two in the bush."

Other dictionaries give an explanation as well as an example of how the idiom is used, cf. esl idiom page (2005):

**keep one's chin up**

remain brave and confident in a difficult situation; don't despair or worry too much. "I know that things have been difficult for you recently, but keep your chin up. Everything will be better soon."

Both dictionaries mentioned here, but also all other phraseological dictionaries that we found on the Internet are characterised by using the mindset so typical of printed dictionaries which only allows searches for the entire lemma, most often on the basis of an alphabetic list of all phrasemes in the dictionary in question. Searches for parts of a phraseme, the meaning of the phraseme or parts thereof, or an example or parts thereof are not possible. The objective here is not to criticize or to make a survey of available phraseological dictionaries which – like other online dictionaries – seem to be of average quality a little or somewhat below that of printed dictionaries.

This article is based on a Danish online dictionary of idioms which was prepared by us, and which has been freely available since 2003: THE DANISH IDIOM DICTIONARY 2003–2005. The dictionary is fairly comprehensive, comprising a total of 8,728 entries, i.e. a dictionary, which in terms of quantity is considerably above the size of most printed idiom dictionaries. In terms of quality, the dictionary is in line with most printed idiom dictionaries and the best online dictionaries, in other words not very impressive. An explanation is given to each idiom, but no examples, no information on the grammatical arguments of the idioms, no indication of synonymous or antonymous idioms, no references to idioms with a similar meaning, in other words no systematic approach. Only a very rudimentary search engine is provided. One can search for the entire idiom, e. g. *gå i gulvet* (*be out for the count*) which results in:

**gå i gulvet**
BETYDNING
besvime eller falde om, fx af overraskelse
ANDRE IDIOMER
gulv
gå

literal translation: **go to the floor**; English idiom**: be out for the count**
MEANING
faint or fall down on the ground or floor, e. g. by surprise
OTHER IDIOM
floor
go

A search can also be performed for part of an idiom, e. g. *dans* ('dance') which gives a total of 8 idioms, including

**gå bag af dansen**
BETYDNING
ikke kunne følge med mere
ANDRE IDIOMER
gå
dans
bag

literal translation: **walk behind the dance**; English phraseme: **fall behind**

MEANING

not being able to keep up the pace

OTHER IDIOMS

walk

dance

behind

The links can also be used for jumping to other idioms, here, for instance, the link *gå* ('walk') which brings one to 166 idioms containing the word *gå* ('walk') or an inflexional form of *gå* ('walk'). Compared to other online idiom dictionaries this is an improvement, albeit a small one. One can search for the entire idiom or parts of it, but not within its meaning. Since there is no systematic way to approach the available meaning items, one cannot search for idioms with a specific partial meaning. If, for instance, one wishes to find an idiom corresponding with 'tired' and 'afterwards', one has no chance of getting to *fall behind* if one does not already know the idiom and can also remember it.

This dictionary was reviewed by Farø (2004), among others. In his review, he points at a number of explanations to a fairly large amount of idioms which are clearly ambiguous or even incorrect. He specifically criticizes the dictionary for lacking grammatical information (valency items) and information on variants, and also the lack of examples featuring the idioms in question. He is also not pleased with the fact that the dictionary does not allow searches of meaning items. Based on his criticism, we have corrected the errors and inaccuracies. Furthermore, his criticism has been a contributive factor in our decision to work out an entirely new concept for a phraseological dictionary, only contributive though, since the most important motivation for the new concept presented in this article are log file analyses of user searches performed in the online dictionary of idioms. We will come back to that later. Farø (2004: 207) quotes a telephone conversation with one of the authors of this article (Henning Bergenholtz) who said the following: "I am not interested in linguistic idiom definitions, only lexicographical ones." To this, Farø objects that he finds it extremely important to work out a really precise idiom definition, both for the sake of the lexicographer who prepares the dictionary as for the dictionary user who will have to do lookups in it. Farø's own definition is extremely sophisticated, but we doubt whether it can be explained in a way that will make the users understand. The following components are included:

- an idiom consists of more than one word
- an idiom is semantically transformed, i.e. what is usually called 'a figurative sense'
- an idiom is semantically speaking neither unequivocal nor precise
- an idiom has an overall meaning, but the meaning of the individual words also come into play

- an idiom has an iconographic function which gives an outline of a metaphoric frame or situation

Farø (2004) considers the definition chosen for THE DANISH IDIOM DICTIONARY to be extremely vague:

> A fixed word combination with the lemma as a core element, which is used only in specific situations and which has a special meaning. As opposed to collocations, idioms are almost fixed in their use and unlike proverbs and familiar quotations an idiom is part of a sentence. An idiom is identified by means of the following rule: If the sum of the meaning of the individual words in a word formation does not correspond with the meaning of the entire word formation, we have en idiom. For example, the phrase *put somebody in the shade* does not mean 'placing'+'somebody'+'in a place without sunshine'. The real meaning is sometimes called the transferred or figurative meaning, in this case 'be very superior to somebody' or 'overshadow somebody'. When, on the other hand, we have a phrase which does not have a clear meaning we do not have an idiom, e. g. *gå fra snøvsen* (*throw a fit* or *lose one's marbles*, since the Danish phrase means both 'be outraged' and 'go crazy'). "Snøvs" in Danish means the upper part of a tied up sack. When tying a sack, you make a "snøvs" by folding the open end of the sack and tying it in a specific way. If the sack is not properly tied, the "snøvs" will slip and the sack will open. It has *lost its "snøvs"* and the contents will be spilt. Here, *snøvsen* does not have a singular clear meaning, nor can it be understood as either the 'mind' or the 'temper'.
>
> If a specific idiom cannot be found in the idiom dictionary, it may be found in THE DANISH INTERNET DICTIONARY as a fixed word formation. According to the definition that we used in order to identify idioms, *gå fra snøvsen (throw a fit* or *lose one's marbles)* is not an idiom. This, on the other hand, is a phrase where the meaning of the entire word formation equals the sum of the meaning of the individual words. (THE DANISH IDIOM DICTIONARY, User Guide)

In our opinion, this definition is rather clear. We understand Farø's (2004) suggestion to search a different dictionary as being helpful to the user, not an admission of failure which may not be what Farø intended. He does say, though, that the definition is quite inadequate.

> Of major theoretical interest to future phraseography is the fact that the idiom concept of the dictionary is neither completely clear to the user nor to the editors themselves which naturally leads to a rather casual lemma selection. The editors, however, are partly aware of the fact that the idiom concept is so vague that it becomes practically difficult to use as a criterion, but to simply make a reference that says "if you cannot find the idiom in one place, try in another" is unsufficient. It must be up to the editors to draw up a suitable and clear idiom concept. (Farø 2004, 233) (our translation RA/HB/VV)

We actually do believe that we have an extremely suitable idiom concept, but we do not believe that all users can or may understand it or know how to use it, cf. Bergenholtz (2006). The question is to which extent dictionary users may be presumed to have the necessary knowledge of linguistic, phraseological or lexicographical theory.

## 2 Phraseology and Lexicography

Strikingly, many, if not all phraseological contributions, are based on available dictionaries or lexicographical needs which the phraseological theories in question are trying their best to solve. Here, in a little unfriendly fashion perhaps, we would like to refer to the well-known maxim *May God defend me from my friends; I can defend myself from my enemies*. The great majority of the phraseological friends of lexicography are characterised by having sophisticated concerns about phraseological classifications and the description of a number of phraseological types. They see it as their mission to help lexicographers (whom they apparently all seem to believe are in dire need of help), e. g.:

> Die Phraseologen sind aufgerufen, den Lexikographen nachvollziehbare Ordnungs- und Einordnungsprinzipien an die Hand zu geben. (Wotjak/Heine 2005: 143)

There is an apparent wish among phraseologists to find the solution to any sort of phraseological classification, once and for all. However, as lexicographers we wish to draw attention to two significant aspects which on one hand do not tell against every sort of phraselogical classification, but on the other hand do nothing to help their general usefulness in lexicography: (1) Lexicography is not a form of applied lexicology or applied phraseology, cf. Wiegand (1989), Tarp (1998) and Bergenholtz (2000: 19ff). (2) Like terminologists, most phraseologists do not distinctly consider dictionaries to be a tool which has been made to help dictionary users cover their needs. It does look as if they believe a certain classification to be relevant for each dictionary type. Here, we are not talking about the many borderline and problem cases discussed in relation with phraseological classifications:

> Nicht selten geht es, bedingt durch bestimmte Randbereichsunschärfen, um Entscheidungen, die ebenso gut zugunsten der einen wie auch zugunsten der anderen Möglichkeit hätten getroffen werden können. Das enthebt die Linguisten und Lexikographen aber nicht der Notwendigkeit Grenzziehungen vorzunehmen und diese zu begründen. (Wotjak/Heine 2005: 144)

Contrary to the theory of a generally valid phraseological classification both in terms of linguistics and lexicography, we wish to emphasize that different dictionary types for different user groups and different user needs most certainly need different classifications. To this thesis must be added, however, that there is a possibility of having different classifications for lexicography and for the user. This criticism of the phraseological literature does not only apply to the article quoted here, but also to the latest phraseological contributions in general, cf. Farø (2005). We do not wish to take a position on the internal phraseological discussion here, but we will say that most contributions are useless to us as lexicographers, both empirically and theoretically, i.e. seen from the point of view of lexicographical theory creation and proposals for new lexicographical concepts for phraseological dictionaries. The main reason is first of all the lack of incorporating and not taking into account the functions of dictionaries, including user prerequisites, i.e. linguistic and cultural previous knowledge and actual needs.

In the modern lexicographical theory of functions, cf. Tarp (1992), Bergenholtz/Tarp (2002, 2003, 2005) the following lexicographical functions are assumed, i.e. functions which a dictionary must comply with in order to meet the needs of the expected users:

communicative functions: text production, reception, translation
cognitive functions: systematic wish for knowledge, sporadic wish for knowledge, ...

By communicative needs is meant help in solving a problem which may arise in relation to reading, producing or translating a text. By cognitive needs is meant help in achieving actual or systematic knowledge of something which may be a linguistic phenomenon or a non-linguistic phenomenon or context. The context of cognitive needs may be illustrated by the following sequence of events which tells nothing about why or in which context the user needs to know more:

dictionary user -> dictionary -> dictionary user

If the dictionary use is communication-related, we have a completely different sequence of events. The user is reading, hearing, correcting, writing or translating a text and he has now come across a specific problem. He does not know how to understand, correct, produce or translate a certain part of the text for which reason he turns to a dictionary for help:

user -> text part$_z$ -> user -> dictionary -> user -> text part$_z$

This does not mean that there cannot be a cognitive gain through a communication-related use of a dictionary; this may be the case, but is not the intention of the dictionary use.

As far as phrasemes are concerned, this division into three basic types of communicative dictionary functions and at least two basic types of cognitive functions can be further divided into several sub-types, all of which could form the basis of a mono-functional phraseological dictionary in its own right. The actual preparation of a fair amount of mono-functional dictionaries is not realistic, and so one must try to compile multi-functional dictionaries which can fulfill several functions at the same time. As for reception, one can make a distinction between meaning items for different user types, e. g. native speakers, non-native speakers, pupils, etc. etc. As for production, one can also distinguish between different user types, but in addition to this and first and foremost between many sorts of production problems, cf. Farø (2004: 205):

- Is there a good idiom connected to this word?
- Is there an idiom to be used in this context?
- Is there a different idiom with almost the same meaning, so that I may vary my usage?
- Is there a different idiom with the opposite meaning of this idiom, so that I may vary my usage?
- Are there any variations which have the same basic stem as this idiom?
- Which collocations does the relevant idiom form part of?

In the case of cognitive functions, it may be a question of curiosity, of acquiring sporadic knowledge of random or specific idioms. This may be in the form of entertainment, the so-called lexicotainment. The other basic cognitive function consists in a wish for new or more systematic knowledge of idioms. For this purpose, consulting the user manual or the theoretical outside matter of a dictionary would be advisable.

When this division is transferred to the phraseological classification problem, it becomes obvious that the user may benefit greatly from and perhaps even wish to gain information on the status of a phraseme, cf. the message of this email which was sent to the authors of THE DANISH INTERNET DICTIONARY:

> Hi
> First of all, thank you for a really great dictionary.
> I have searched your dictionary repeatedly trying to find out whether the phrase "dreje sig om" (literal translation: "turn about"; English translation: 'be about') is an idiom or a collocation. Can you possibly help me with this?
> Best regards
> NN

To this question we gave the following answer:

> According to our standard definition of idioms "dreje sig om" ('be about') is a collocation. The explanation can be found under "dreje" ('turn') in the sub-article "dreje sig om" ('be about'):
>
> **dreje**
> **...**
> 5. **dreje sig om**
> BET være et spørgsmål om
> = anbelange; angå; berøre; gælde; handle om; vedrøre; være tale om
> KOL dreje sig om nogen; dreje sig om noget
> EKS Hvad drejer det sig om?; Han tror, alting drejer sig om ham.; Det drejer sig om minutter.
>
> **turn**
> **...**
> 5. **be about**
> MEANING be a question of
> = have to do with; concern; be relevant to; refer, pertain, relate to; deal with
> COLLOCATION be about somebody; be about something
> EXAMPLE What is this about?; He thinks everything is about him.; It is a question of minutes.
>
> An idiom consists of at least two individual words that each have their own clear meaning (when they appear individually). This is not the case with "dreje sig om" ('be about'), but it is true that the entire phrase "dreje sig om" ('be about') must be explained as a whole. Therefore, a different definition type than the one we have chosen might bring about a different classification. Cf. the user manual of THE DANISH IDIOM DICTIONARY.

> But what is an idiom really? In THE DANISH INTERNET DICTIONARY the following explanation is briefly mentioned: "fixed word formation with a special meaning". It is this meaning which forms the basis of the distinction between normal word formations (called collocations) and special word formations (called idioms). To these must be added proverbs, sayings and quotations which differ from idioms by always making up an entire sentence. It may also be said that an idiom is a word formation in which the meaning of the entire word formation does not correspond with the sum of the meaning of the individual words. We have used this methodical basic consideration, acting on the meaning items and the synonyms in THE DANISH INTERNET DICTIONARY.
>
> The distinction between the broader concept of 'fixed word formation' and idiom is not entirely clear. But according to the definition that we have used, *gå fra snøvsen* (*throw a fit* or *lose one's marbles)* is not an idiom. What we have here is a phrase where the meaning of the entire phrase may be understood as the sum of the meaning of the individual words (since the meaning of *snøvsen* is unclear, but must be seen as something like the 'mind' or the 'temper'). If you have any doubts, you should also check THE DANISH INTERNET DICTIONARY.
>
> Yours sincerely,
> Centre for Lexicography

Despite every effort to help the users we may conclude, based on a log file analysis of the use of THE DANISH INTERNET DICIONARY and THE DANISH IDIOM DICTIONARY, that a classificatory question from the users like the one quoted above is an absolute exception. The typical question will often refer to an assumed lemma lacuna, cf. the following email from a user of THE DANISH IDIOM DICTIONARY:

> I couldn't find the phrase 'Lige børn leger bedst' (literal translation: "Equal children play the best", English translation: *Like will to like*) in your dictionary. That must be a mistake, surely?

Our answer was:

> The question is: "What is an idiom?" and "What is a proverb?"
> *Lige børn leger bedst (Like will to like)* is a proverb according to our definition. In the user manual you will find a definition of both terms. The phrase can be found in THE DANISH INTERNET DICIONARY as a proverb both under the entry "barn" ('child') and under "lige" ('equal').

We have given similar answers to a number of similar questions like for instance *gøre sig selv en bjørnetjeneste* (literal translation: "do oneself a bear's service", English idiom: *cut off one's nose to spite one's face)* and *gå fra snøvsen (throw a fit* or *lose one's marbles)*. These are collocations which one may find in a different internet dictionary. We will maintain that this is not a bad answer, for it does apply the methodical basis for the distinction made by the lexicographer of idioms on the one hand and collocations, proverbs and quotations on the other (which, naturally, are not seen as one and the same phraseme type, but as different ones). Unlike collocations, idioms may not be understood as the sum of individual meanings of all the words of the phraseme, and therefore, it must be explained as a whole. In the case of a collocation, one only has to know the meaning of each individual word. Then the user can find the explanation in a major common language dictionary – for every single word of the collocation. When a common language dictionary contains a meaning item for *bjørnetjeneste*

which also corresponds with the use of this word in the phraseme *gøre sig selv en bjørnetjeneste* (*cut off one's nose to spite one's face*), we have a collocation and not an idiom. Do the users understand this definition, however? The answer is yes – to a certain degree. Within the first year of the idiom dictionary being in existence, two out of three searches made by the users resulted in the desired idiom (the number of users for 2005 is more than three times as high):

| **Use of The Danish Idiom Dictionary** | | |
|---|---|---|
| **Number of lookups** | **42,749** | |
| stating one or more idioms | 30,085 | (70.4%) |
| string not found | 12,664 | (29.6%) |

With a share of 30% of the lookups, the number resulting in strings not found is somewhat larger than the corresponding number in the common language dictionary, where the following numbers represent the log files of the exact same period of time:

| **Use of The Danish Internet Dictionary** | | |
|---|---|---|
| **Number of lookups** | **1,016,960** | |
| stating one or more dictionary entries | 818,613 | (80.5%) |
| string not found | 198,347 | (19.5%) |

In this common language dictionary (THE DANISH INTERNET DICTIONARY) about half of the lookups that resulted in strings not found are caused by lemma lacunas (the entry searched for is not found in the dictionary), and the other half of the lookups resulting in strings not found are caused by spelling errors (for more detailed information, cf. Bergenholtz/Johnsen 2005). In THE DANISH IDIOM DICTIONARY things are quite different. The number of spelling errors is more limited, less than 3% of all lookups. Searches for real idioms resulting in strings not found are also much more rare than searches for entry words in THE DANISH INTERNET DICTIONARY, probably because it is an idiom dictionary with a particularly large number of lemmas – more than 8,000 idioms which is more than twice the size of most idiom dictionaries. In less than 1% of the cases, the search was made for a lemma lacuna. Here, most searches are made for non-lemmatised variants such as *vogte sit skind* (literal translation: "watch one's skin"; English idiom: *save one's bacon*), which could not be found in the dictionary, even though the variants *hytte sit skind* (literal translation: "look after one's skin"; English idiom: *save one's bacon*) and *vare sit skind* (literal translation: "take care of one's skin"; English idiom: *save one's bacon*) were found. Most searches resulting in strings not found were cases, in which the user searched for a phraseme which was not an idiom. In this case, the log files of the idiom dictionary show that afterwards many users find the desired phraseme in the common language dictionary, THE DANISH INTERNET DICTIONARY. The result is not satisfactory, though. If more than 25% of all lookups in an idiom dictionary involve searches for phrasemes which are not idioms, our conclusion is not that we must find a new and better definition of the idiom concept. Instead, the dictionary must take into account the fact that the user, when needing help with text reception and text production, i.e. with communicative dictionary functions, wishes to get information on a phraseme, but he is not particularly interested in the status of this phraseme. When it comes to cognitive needs –

also known as knowledge-related needs, it is a different case. Here, the gaining of information on the classificatory status of the phrasemes is important to the user.

## 3 Conception of an Entirely New Form of Phraseological Online Dictionary

Most idiomatic printed dictionaries are polyfunctional, but in practise an electronic dictionary can be used as a monofunctional dictionary. Here, one can predict that the user will define his user situation by choosing one of the three following user types:

1. I need help when reading a text
2. I need help when writing a text
3. I want to know more about idioms

According to choice, the user will get different types of information when using the dictionary (more in the last chapter).

The dictionary includes 2,662 collocations which in our opinion will need to be explained, e. g. *gøre sig selv en bjørnetjeneste* (*cut off one's nose to spite one's face*), 1,472 proverbs and 8,728 idioms. The lexicographer must fill in most of the fields himself; others will automatically be added by the program (synonyms). A typical article will have the following filled-in fields:

1. **a field stating the core of the phraseme, i.e. lemma**
2. **a field stating the type of phraseme**
   (idiom, collocation, proverb)
3. **a field stating style level**
   (high, neutral, low. If the lexicographer does nothing, the style will automatically be stated as neutral)
4. **meaning item/explanation**
5. **grammatical collocation**
   (by this is meant valency items using keywords)
6. **note**
   (here, etymological information and perhaps encyclopedic information which goes beyond the meaning item)
7. **internet address**
   (if there is a text about or with special relevance to this particular phraseme, a reference to this can be made here)
8. **collocation**
   (here, collocations which include the phraseme may be stated)

9. **example**

   (here, an example which includes the phraseme may be stated)

10. **synonym**

    (these synonyms are made automatically by the program if two phrasemes have the same meaning item)

11. **antonym**

    (here, phrasemes with a meaning item which can be called 'opposite' may be stated)

12. **reference**

    (here, a phraseme which has a related meaning without being a synonym or an antonym may be stated)

13. **associative words**

    (here, words that the lexicographer associates with the phraseme may be stated; the users themselves may add further associative words here)

The following example shows how the systematics are applied.

**kaste med mudder**
FRASEMTYPE idiom
STIL neutral
BETYDNING
nedgøre ved at komme med nedsættende eller kritiserende bemærkninger
GRAMMATISK KOLLOKATION
nogen kaster med mudder efter nogen
nogen kaster med mudder efter noget
nogen kaster med mudder mod nogen
nogen kaster med mudder mod noget
KOLLOKATION
aldrig kaste med mudder
begynde at kaste med mudder
være hurtig til at kaste med mudder
EKSEMPEL
Men uansat hvad skal du være varsom med at kaste med mudder. Selv om du mest har lyst til at svine kunden til, skal du tænke på, at det kan skade dit firma.
SYNONYM
drage ned i mudderet
drage ned i smudset
hænge ud
ikke levne for to skilling ære
kaste skarn på
kaste skidt på

kaste smuds på
kaste snavs på
køre på
pege fingre ad
rakke ned
sable ned
smide med mudder
sudle til
svine til
sværte til
søle nogens navn til
trække ned i mudderet
trække ned i smudset

ASSOCIATIONSORD
tilsvine
angreb
angribe
bagtale
mudderkastning

**sling mud**

PHRASEME TYPE idiom
STYLE neutral
MEANING
damage somebody's reputation by criticizing or saying bad things about them
GRAMMATICAL COLLOCATION
somebody slings mud at somebody
somebody slings mud at something
somebody slings mud against somebody
somebody slings mud against something
COLLOCATION
never sling mud
start slinging mud
be quick to sling mud
EXAMPLE
But no matter what, you have to be careful when slinging mud. Even though you would like nothing more than to say bad things about the client, you must remember that it may harm your company.
SYNONYM
drag down into the mud
drag down into the dirt
blame
not leave any honour
fling mud at
throw filth on

besmirch somebody's good name
drag somebody's name through the mud
throw dirt on
point the finger of scorn at
put down
run down
throw mud at
defame
blacken
smear somebody's name
pull down into the mud
pull down into the dirt
ASSOCIATIVE WORDS
smear
attack (noun)
attack (verb)
slander
mudslinging

## 4 Searching the Dictionary

Basically, an electronic dictionary consists of two parts: the search interface and the result. For a search to be successful, the data to be searched needs to be properly structured. In this case, the intended technical possibilities of the database and the implications of its design had an influence larger than usual on lexicographical decisions in that the idiom, the main type of data found in the dictionary, is entered in an amputated form. For instance, *anmode om nogens hånd* ('ask for somebody's hand in marriage') is reduced to *anmode om hånd* ('ask for hand in marriage'). The full idiom is then entered into grammar fields, example fields, etc. The reason for this is that few, if any, users will enter a search string that exactly matches the corresponding idiom in the database. The word *nogens* ('somebody's') might as well be entered as *pigens* ('the girl's'), and a basic search will return a "string not found" error message which in this case would be right from a technical point of view and wrong from a lexicographical point of view. Hence the decision to amputate idioms into core strings. The following explains how this solved a crucial problem: to guess what the user is looking for.

Basic searches like the ones found in THE DANISH INTERNET DICTIONARY that only searches in one specific location in the database are of little value when considering the user needs in an idiom dictionary like the proposed DICTIONARY OF DANISH PHRASEMES. This dictionary must provide the user with optional search tools. The current solution proposes to search both for strings that are related directly, i.e. based on orthography similar to traditional dictionary searches, and indirectly based on context, i.e. based on association.

A search based on orthography in the dictionary will take the following path:

1. The user enters his search string into the designated field and starts the search. The search string can be of any length containing any number of "words".

2. The software will try to find a matching string in the lemma field, where idioms, collocations, and proverbs are stored, in all the records of the database. At this point, most search engines stop and the user will either be presented with all the found records, be asked to enter a different search string in case no records were found, or refine his search to reduce the number of found instances. As follows, however, this may be done partly by the software instead.
3. The search string is parsed by the program into its constituent units/"words".
4. A number of searches are performed whereby the original search string is reduced to its parts and combinations hereof in order to find any record that contains the maximum number of units/"words" entered in the search string. I.e., the words entered by the user are re-ordered to find any combination that matches the maximum combination possible of similar words in the lemma field of the dictionary. The result may not be what the user was looking for, but he will at least have a result and can deduce the software's overall function.

The prime concern is to find a method that allows a result even when the user is entering a string that only on very rare occasions completely matches the lemma field in the database. For example: the string *anholde om en piges hånd* ('ask for a girl's hand in marriage') would be impossible to find in the database, since it is not lemmatised because the string *en piges* ('a girl's') is replaceable by a name or similar attribute. What is lemmatised, however, is *anholde om hånd* ('ask for hand in marriage'). A search for *anholde om en piges hånd* ('ask for a girl's hand in marriage') would therefore be futile, which the user has no chance of knowing. Traditionally, the user would be asked to repeat/refine/relinquish his search without further ado. Such a strategy is pure laziness on the part of the programmer. It is relatively simple to let the software do some of the work a user would be forced to do by hand. Our proposal is to search for each unit/"word" by itself in the order of which it was entered and to try to eliminate replaceable parts through repeated searches following a fixed pattern:

ISSN 1615-3014

| | **Search (algorithm)** | **Result** |
|---|---|---|
| **Step 1** | Search for any string that starts with the string *anholde* | *anholde om hand* |
| **Step 2** | Is the string *anholde om* present in the found set? | Yes |
| **Step 3** | Is the string *anholde om pigens* present in the found set? | No |
| **Step 4** | Remove *pigens*. Is the string *anholde om hånd* present in the found set? | Yes |
| **Step 5** | Show article | |

This is a very crude way of helping to find a given string and the method will be developed further, but it shows the principle at hand. The lexicographer should be able to foresee a certain behaviour on the side of the user in order to anticipate his next move when presented with the problem of finding a given string in a dictionary. Further refinement is possible and could even be elaborated upon by including the user's participation creating a "system-assisted search". In the dictionary at hand that would be a bit excessive, though.

The above example is very simple since at current there are no other lemmas that contain the word *anholde* ('ask for'). The word *hånd* ('hand'), however, is much better represented. A search for that string results in 86 records. The point is that a standard search like the ones performed in THE DANISH INTERNET DICTIONARY and THE DANISH ACCOUNTING DICTIONARIES that require precise partial or complete orthographic matches allowing only for grammatical variations, is too rigid to be of any real value when trying to find a particular idiom. Also, adding to the dictionary all possible word forms and variations of a given idiom is a never ending task. A basic search will remain being the starting point of the search since, in principle, one could be lucky and the precise string entered is actually in the database which would speed up a search considerably.

More often than not, this is not the case and automatic refinement will take place. The drawback is speed, or the lack of it. Depending on hardware and bandwidth, this potentially complex method causes delays that the user may find counter-productive. How long these delays are going to be cannot be assessed at current, since no live prototype exists on the Web on which to conduct tests. However, programming both basic and extended searches and allowing the user to choose which ones to use is a small matter and could easily be implemented. Nevertheless, this should not be done until tests have been conducted on the speed penalty that a more complex search method implies. The not so apparent advantage for the user is that although the search slows down, the extended search has been programmed by someone who knows the precise conditions of how the data is entered and his algorithm, although possibly overly precise, has better chances of finding anything quickly than the user has through repeatedly entering variations of a search string.

The second search method, search by context or association, is more difficult to describe. Calling the method associative or contextual is not quite precise but it is the best term we can give the method at current.

A given idiom can be seen as an image. Each word alone has a given meaning but together they form a different one. For instance, the idiom *have en kort lunte* ("have a short fuse") has nothing to do with explosives. Nevertheless, the idiom provides an image of a person's temper by liking him to a bomb that is easily set off. Idioms are similar to images and describing them causes the same type of problems that image-databases have tried to solve for over a decade.

Image-databases, specialised databases for multimedia data, have dealt with the problem of how to find a particular image in various ways. Amongst the methods tried have been attempts to analyse an image for its parts by pattern recognition in order to automatically classify images into orders of *people*, *mountain*, *clouds*, etc. None of these often semi-automatic methods proved valid. Two manual methods, however, have proven moderately successful. In one, the administrator of the database builds a list of topics. The user who imports and edits the images is then given the task of applying this limited set of keywords to the images he imports into the database. The problem is that when he imports an image for which there is no keyword he has to ask the administrator to create one. The user, on the other hand, has to know the precise terminology in the database and usually consulted the list before searching for an image. The method is cumbersome and has since been relinquished in favour of the alternative method.

The alternative to rigid topic lists is that each user enters the keywords that he finds to be appropriate as needed, in other words, keywords by association. This data is often supplemented with metadata such as in which publication or article the material has been used, date- and timestamps, photographer's name, copyright information, etc. The method may seem haphazard and confusing, but has been quite successful since its introduction. This in part because image-databases are primarily in use amongst photographers and PR-houses, i.e. among creative users who find rigid lists impractical. The primary advantage in open, uncensored lists is that you never run out of keywords and each addition to the list makes it easier to find a given image.

In the dictionary of idioms the same method will be used to add keywords by association. To begin with, the editor enters a maximum of five associative keywords that he finds appropriate for the idiom in question. Preferably, these keywords are entered intuitively without long considerations. The essence here is not which keyword best describes the idiom, but what images the idiom brings into the editor's mind. Needless to say, different editors will supply different keywords at different times. Unlike images taken with a camera where keywords are based on either or both motif (*water, summer, beach, people*) or emotions (*happy, relaxed*), idioms provoke purely emotional keywords along the lines of how and what you feel by reading it. However, since the idioms are taken out of context they can provoke a number of associations depending on who and when they are read, simply because each reader will supply his own context, a context that might vary from day to day. For an idiom

this is an enrichment. Taking the above idiom *have en kort lunte* ("have a short fuse"), one person might describe it with the Danish equivalents of the words *temper*, *bomb*, *arguing*, while another might add *taxi driver*, *discussions*, and so on.

The use of associative keywords to assist a search is meant as a help in a text production situation. A user who is in the middle of writing a speech may at some point wish to use an idiom. Maybe he knows it already either partially or in context, but he is not sure of its proper orthography or meaning. In this case a basic search is enough. But sometimes a user wishes to find an idiom with a certain meaning that fits the context of his text. So far, users looking for help here have been let down. The option to search for a number of keywords, optionally combined with a basic search can result in a number of idioms that share common ground. Searching for the keyword *død* ('dead') in order to find all idioms concerning such a grave matter would result in a great number of idioms none of which contain any orthographic form of the word *død* ('dead'), which, by the way, is the very point of an idiom.

So far, the number of associative keywords per idiom is limited to five. This is a random limit and five keywords per idiom are far from enough to be of any practical value. It would, however, be a herculean task for the editor to compile a sizeable list. Even five keywords instead of the originally proposed three add a considerable time penalty in the editing process. We are therefore considering an increase of the number of keywords in the database through user interaction. This could be done in a simple way, after all, all search strings that are being entered will be logged, evaluated, and some of them added to the database by the editor. Why not let the user himself add to the list every time he finds an idiom? Over time, the amount of keywords will increase which should facilitate its use for other users. This would of course be optional, but it is quite possible that many users might find it rewarding to add to a dictionary in this way, which the Wiki-projects clearly prove. Not all these keywords will be equally helpful, but the point is to combine keywords when conducting a search.

For example: a large number of idioms are on the subject of responsibility, how to obtain it, and how to relinquish it. They all share the associative keywords *ansvar* ('responsibility').

Suppose the idioms *få aben* (literal translation: "be given the monkey"*)*, *sidde tilbage med aben* ("be left with the monkey"), and *have aben på sin skulder* ("have the monkey on one's shoulder") - all three meaning 'be given an undesirable or disadvantageous responsibility' - are all labelled with the associative keywords *ubehagelig* ('unpleasant') and *embedsmand* ('civil servant'). The user searches for an idiom with the following parameters:

> The idiom must contain the word *abe* ('monkey')
> Associative keywords required are: *ansvar* ('responsibility'), *embedsmand* ('civil servant') and *frasige* ('relinquish').

The search would then take on the following steps:

| | **Search (algorithm)** | **Result** |
|---|---|---|
| **Step 1** | Search for idiom containing any form of the string *abe* that also contains the keyword *ansvar* | 3 found records |
| **Step 2** | Select from these 3 those with keywords *ansvar* AND *embedsmand* | 3 found records |
| **Step 3** | Select from these 3 those with keywords *ansvar* AND *embedsmand* AND *frasige* | 0 found records |
| **Step 4** | Return to last found set of idioms | |
| **Step 5** | Tell the user that his search has given partial results. List the 3 idioms in step 2 | |

As an option, it should be possible to choose other parameters to help refine the search, for instance "Show all records where 2 or more of the associative keywords are present" where "2" is a user choice.

The results of the example may seem as if the algorithm has been less than helpful. In reality two very important results have come from the search:

First, there are no idioms in the database that relate to monkeys **and** relinquish, which should trigger the notion that monkey-idioms are all about responsibilities taken and/or given and second, three idioms, however, do contain some of the associative keywords (list supplied) and these are all about taking and/or giving responsibilities.

Further, the user is given two hints the purpose of which is to tell the user what his result is on associative keywords alone:

- Each idiom is supplied with a list of matching associative keywords. After each, a number in parentheses reveals how many other idioms are associated with the same keyword. A click on the keyword will list these other idioms.
- A list of idioms containing all of these three associative keywords regardless of other parameters entered by the user. It would be preferable if this were a list of all possible combinations, but in cases of more than five keywords the number of possible combinations could pose serious limitations depending on hard- and software. It is worth looking into the possibility, though.

The notion of randomly assigning associative keywords to idioms has been criticized for not being sufficiently rigorous and for being impossible to manage. Linguists would prefer to have closed lists of systematic keywords that clearly classify idioms into a fixed terminology, like PONS WÖRTERBUCH - DEUTSCHE IDIOMATIK (Schemann 1993) has done. The opposite, a randomly created and disorganised list, disturbs the notion of academia requiring lists to be managed, controlled, and systematised. The question is whether managing a list, any list, is even necessary when publishing any given database. The editor will without a doubt feel a

need to check and refine the lists, but this is in no way necessary because of a peculiarity that distinguishes online dictionaries from printed ones:

*a word does not exist in a database until it is being searched for*

…metaphorically speaking, that is. The opposite of this is also interesting. What if a word is searched for that does not exist in the database? Is the database then in fact non-existent? Is the terminology that classifies it? The idea that words do not exist before actually spoken or searched for is the same reason why online dictionaries have no upper boundary for lemmata, nor need to constrain themselves to words from a given terminology. Thus, a fixed list of keywords can never be correct no matter how extensive. This has been tried before, and because some people never learn, they still try. There are not enough words to describe all words. Please stop trying! Apart from the futility of creating a fixed terminology for idioms let alone classifying them, a closed list would prove counter-intuitive to the user and hence would not be used. He would simply give up trying to understand the system and search on a best-guess-basis. There is, however, a weakness inherent in the proposed system of disorganised lists. Others have failed to mention it or even see it, but it exists.

The main weakness of the system is that the user will have expectations that cannot be met. No editor can ever expect to guess the way others think or associate. The axiom "there are not enough words to describe all words" works against this system, too. Combining monkeys with bureaucrats is a very subjective notion and it is doubtful that the user would have had the same train of thought as the editor. However, it is not a question of finding the "right" or "wrong" associative keywords. In the end, it is a simple question of resources: do we want to finish the dictionary or not? Five keywords are enough to start with. Find them quickly, do not try any harder and focus on the important part: finish the dictionary! In time, categorisation will take place, but not from a fixed terminology.

There are two main approaches to adding associative keywords: an intuitive and an academic one or in other words association and categorisation. The editor can work his way through the database line by line intuitively associating one with the next creating synonyms and antonyms as he finds them and adding keywords to groups of idioms. This type of categorisation is no worse or better than most cataloguing of items. Definitions that contain the word *ansvar* (*responsibility*) will most likely all end up with the keyword *ansvar* ('responsibility') or a variation, just as all idioms symbolising death will contain *død* ('dead') as a keyword. Hence, a certain categorisation **is** taking place. Furthermore, having THE DANISH INTERNET DICTIONARY at our disposal makes it a given to try and match both synonyms and antonyms to keywords so that a user searching for *afdød* ('deceased') will automatically be led to idioms marked *død* ('dead'). Technically, this will be part of a "last resort search" and has to be considered very carefully. Including synonyms and synonyms of synonyms in a search can cause serious congestion in any database system because every search will be repeated continuously, and in the worst case perpetually.

User-added associative keywords are important because they have the possibility of being truly associative. Users can never see more than a few of all idioms recorded. For them, all keywords are given to ungrouped idioms. They may expect categorisation like *ansvar* ('responsibility') and *død* ('dead'), but they will not and cannot by themselves without proper insight into the database be able to categorise large groups of idioms. And even if they did, the classification might only be of use to themselves since no classifications or keywords are listed openly anywhere in the dictionary.

Naturally, a certain amount of sabotage including addition of senseless associative keywords must be expected, but such activity will have to be planned quite carefully on the part of the saboteur. Nevertheless, to discourage this type of activity a method will be implemented that requires users to accept responsibility. This could be done simply by requiring an email address to which a conformation is sent. If the email is rejected by the attended recipient the submitted form is neglected.

Although it is important to let the user supplement the dictionary with associative keywords it is more important to let him provide new idioms or other dictionary entries. This will be accomplished through a standard input form with fields for idiom, description, examples, etc. The user will submit this form and the data will be reviewed and edited by the editor in case the idiom gets accepted.

The advantage of this possibility is not to be neglected. There are two primary gains by involving the user in the process of editing the dictionary. First, as users of THE DANISH INTERNET DICTIONARY have proven, many users feel a need to enrich a dictionary with their own findings thus creating a solid and loyal user base. Such a user base might even recommend the dictionary to other potential users. In part, they are endorsing their own work further legitimising the dictionary's content. This may seem cynical, but the dictionary gains more by user involvement which leads to the other prime reason for doing so. Second, no matter how vigilant the editor may be and how many hours he spends editing the dictionary many idioms and variations thereof will not be found. Users producing a text find idioms in other contexts simply because they are missing them. Consider the opposite problem, reception of a text, where the user cannot find an idiom in the dictionary and wishes to supplement or vent grievances about this apparent lack. The concept has been tested successfully in THE DANISH INTERNET DICTIONARY more by accident than by thought, where users often submitted words for entry into the dictionary simply because they could not find them. These words are gifts and must be taken seriously and again nothing is lost nor is the quality of the dictionary diminished by doing so, since a word not searched for does not exist. Therefore, even if the entry only exists in the user's mind and only will be searched and found once, no harm is done. Even if it is a phrase that some may consider slang.

## 5 Results

A user will be presented with different results depending on how he predefines his user situation. He will be given the choice between three possibilities, either of which will

influence the amount of data that he will be presented with. For instance, in a given situation the user defines any or all of the following search criteria:

- search for the lemma *kaste med mudder* ("sling mud")
- search for parts of the meaning item 'damage somebody's reputation by criticizing or saying bad things about them'
- search for the associative keywords *slander, mudslinging, attack*

If the user defines himself as **I need help when reading a text,** only the lemma with its meaning item will be shown:

**sling mud**
MEANING
damage somebody's reputation by criticizing or saying bad things about them

If he defines himself as **I need help when writing a text,** he will be presented with every piece of information that relates to the lemma *kaste med mudder* ("sling mud") with the exception of remarks and the list of associative words.

**sling mud**
PHRASEME TYPE idiom
STYLE neutral
MEANING
damage somebody's reputation by criticizing or saying bad things about them
GRAMMATICAL COLLOCATION
somebody slings mud at somebody
somebody slings mud at something
somebody slings mud against somebody
somebody slings mud against something
COLLOCATION
never sling mud
start slinging mud
be quick to sling mud
EXAMPLE
But no matter what, you have to be careful when slinging mud. Even though you would like nothing more than to say bad things about the client, you must remember that it may harm your company.
SYNONYM
drag down into the mud
drag down into the dirt
blame
not leave any honour
fling mud at
throw filth on
besmirch somebody's good name
drag somebody's name through the mud
throw dirt on
point the finger of scorn at

put down
run down
throw mud at
defame
blacken
smear somebody's name
pull down into the mud
pull down into the dirt

If he defines himself as **I want to know more about idioms,** he will be presented with every piece of information that relates to the lemma *kaste med mudder* ("sling mud"). In this case all of the above plus commentaries, if any, and all associative keywords:

**sling mud**
PHRASEME TYPE idiom
STYLE neutral
MEANING
damage somebody's reputation by criticizing or saying bad things about them
GRAMMATICAL COLLOCATION
somebody slings mud at somebody
somebody …
COLLOCATION
never sling mud
start …
EXAMPLE
But no matter what, you have to be careful when slinging mud. Even though you would like nothing more than to say bad things about the client, you must remember that it may harm your company.
SYNONYM
drag down into the mud
drag …
ASSOCIATIVE KEYWORDS
smear
attack (noun)
attack (verb)
slander
mudslinging

We plan to finish the new online phraseological dictionary by mid 2007. It will be called DICTIONARY OF DANISH PHRASEMES.

## References

Bergenholtz, Henning (2000): "Lexikographie und Wortbildungsforschung". In: Barz, Irmhild/Fix, Ula (eds.): *Praxis- und Integrationsfelder der Wortbildungsforschung,* Heidelberg: 19–30.

Bergenholtz, Henning (2006): "Idiomwörterbücher und ihre Benutzer". In: Breuer, Ulrich/Hyvärinen, Irma (eds.): *Wörter – Verbindungen. Festschrift für Jarmo Korhonen*, Frankfurt a.M. etc.: 19–30.

Bergenholtz, Henning/Johnsen, Mia (2005): "Log files as a tool for improving Internet dictionaries". *Hermes* 34: 117–141.

Bergenholtz, Henning/Tarp, Sven (2002): "Die moderne lexikographische Funktionslehre. Diskussionsbeitrag zu neuen und alten Paradigmen, die Wörterbücher als Gebrauchsgegenstände verstehen". *Lexicographica* 21: 253–263.

Bergenholtz, Henning/Tarp, Sven (2003): "Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions". *Hermes* 31: 171–196.

Bergenholtz, Henning/Tarp, Sven (2005): "Wörterbuchfunktionen". In: Barz, Irmhild/Bergenholtz, Henning/Korhonen, Jarmo (eds.): *Schreiben, Verstehen, Übersetzen und Lernen: Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*. Frankfurt a.M. etc.: 11–25.

DICTIONARY OF DANISH PHRASEMES = Bergenholtz, Henning /Vrang, Vibeke: *Ordbogen over faste vendinger*. Database concept and design: Richard Almind. Århus. www.idiomordbogen.dk/ (in preparation).

ESL IDIOM PAGE (2005): *ESL Idiom page*. www.pacificnet/~sperling/idioms.cgi. (November 2005).

Farø, Ken (2004): "Idiomer på nettet: Den danske idiomordbog og fraseografien". *Hermes* 32: 201–235.

Farø, Ken (ed.) (2005): *Fraseologi. Thematic Section. Hermes* 35: 11–181.

GOENGLISH.COM (2005): *GoEnglish. com*. www.goenglish.com/Idioms.asp. (November 2005).

Tarp, Sven (1992): *Prolegomena til teknisk ordbog*. PhD dissertation. Aarhus: Department of Spanish. Aarhus School of Business. www.lng.hha.dk/dml/spa/phd.pdf. (November 2005).

Tarp, Sven (1998): "Leksikografi på egne ben. Fordelingsstrukturer og byggedele i et brugerorienteret perspektiv". *Hermes* 21: 121–137.

THE DANISH IDIOM DICTIONARY (2003–2005) = Vibeke Vrang with contributions by Henning Bergenholtz and Lena Lund: *Den danske Idiomordbog*. Database and design: Richard Almind. www.idiomordbogen.dk. (November 2005).

THE DANISH INTERNET DICTIONARY (2002–2005) = Henning Bergenholtz with contributions by Vibeke Vrang, Lena Lund, Helle Grønborg, Maria Bruun Jensen, Signe Rixen Larsen, Rikke Refslund and Mia Johnsen: *Den Danske Netordbog*. Database and design: Richard Almind.www.dendanskenetordbog.dk. (August 2005).

Wiegand, Herbert Ernst (1989): "Der gegenwärtige Status der Lexikographie und ihr Verhältnis zu anderen Disziplinen". In: Hausmann, Franz Josef et al. (eds.): *Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. An International Encyclopedia of Lexicography. Encyclopédie internationale de lexicographie. Erster Teilband*. Berlin/New York: 246–280.

Wiegand, Herbert Ernst (2005): "Äquivalentpräsentation und Wörterbuchfunktion in zweisprachigen Printwörterbüchern. Mit einem Seitenblick auf die so genannte "moderne lexikographische Funktionslehre"". *Germanistische Linguistik* 179: 1–38.

Wotjak, Barbara/Heine, Antje (2005): "Zur Abgrenzung und Beschreibung verbonominaler Wortverbindungen (Wortidiome, Funktionsverbgefüge, Kollokationen)". *Deutsch als Fremdsprache* 42: 143–153.